ACEEE

Int. J. on Recent Trends in Engineering and Technology, Vol. 10, No. 2, Jan 2014

A Color Boosted Local Feature Extraction Method for Mobile Product Search

Kaiman Zeng, Nansong Wu, and Kang K. Yen
Florida International University/Department of Electrical and Computer Engineering, Miami, USA
Email: {kzeng001, nwu, yenk}@fiu.edu

***Abstract*— Mobile visual search is one popular and promising research area for product search and image retrieval. We present a novel color boosted local feature extraction method based on the SIFT descriptor, which not only maintains robustness and repeatability to certain imaging condition variation, but also retains the salient color and local pattern of the apparel products. The experiments demonstrate the effectiveness of our approach, and show that the proposed method outperforms those available methods on all tested retrieval rates.**

***Index Terms*— SIFT, color SIFT, feature extraction, mobile product search**

I. INTRODUCTION

In recent years online shopping has become an important part of people’s daily life with the rapid development of e-commerce. As a new fashion of shopping, online shopping has surpassed or even replaced the traditional shopping method in some domains such as books, CD/DVDs, etc. Product search engine, as an imperative part of the online shopping system, greatly facilitates users retrieving products and doing shopping. The traditional product search is based on keyword or category browser. Since in e-commerce products are usually presented in the form of images, the visual search, which retrieves an image by its visual contents and features, brings a new insight into product search and attracts growing interests among researchers and practitioners. In particular, mobile visual search is one of the promising areas because of the increasing popularity and capability of mobile devices.

Traditional content-based image retrieval methods adopt a model learnt from text information retrieval. The features of image are treated as visual words, and the image is represented by Bag of Words or Bag of Features. A typical mobile product search process is that a user takes a picture of a product with a camera phone, the visual features of the product is extracted, the correspondence between features of query image and database images is evaluated, then the product is identified and/or its visually similar products are retrieved from the database. The search results greatly depend on the detected feature used and may vary tremendously. Color, shape, texture, and local feature are some of the most common features in search.

In this paper, in respect to the characteristics of product image especially the apparel product image we propose a novel feature extraction method, which combines product color feature and local pattern feature in a way that they complement each other. For apparel products the color, texture and styles are sometimes difficult or unclear to express in words, while the images provide a good and natural source to describe thesefeatures. Difference from generic image search, there must be an interested object aligned in the center of the imagein product search. On the other hand, the imaging position will result light, scale, and affine variation. Hence, this paper will focus on developing a local feature descriptor that not only maintains robustness andrepeatability to certain imaging condition variation, but also retains the salient color and style

DOI: 01.IJRTET.10.2.37

features of the apparel products. We break down the problem of identifying an apparel item from a query image into threemain stages: (i) extract the item features from a color image, (ii) match its visual features to a large imagedataset of apparel items, and (iii) return a set of matching images with its brand and style information.We match apparelitems in images by combining two visual features of color and local features in a complementary way. The paper is organized as following outline. In Section II, we review feature extractions including keypoint detection, local feature descriptors, and color feature descriptors. In Section III, we discuss the system framework and the design of proposed color boosted local feature descriptors. In Section IV, we perform several experiments on an apparel dataset and summarize the results. Finally, we make a conclusion and propose future working areas in Section V.

II. Literature Review

Many mobile image search applications have emerged such as Google "Goggles", Amazon "Snaptell", IQ Engines "oMoby", Stanford Product Search system etc. Most of them employ the local feature-based techniques [1]. Feature extraction phase is typically divided into two stages - keypoint detection and description of local patches. Harris corners [2] is a classic keypoint detection algorithm. Mikolajczyk [3] takes the scale space theory into consideration and proposes Harris-Laplace detector, which applies Laplace-of-Gaussian (LoG) for automatic scale selection. It obtains scale and shape information and can represent local structure of an image. Lowe [4] applies Difference-of-Gaussian (DoG) filter, an approximate to LoG, in SIFT algorithm to reduce the computational complexity. Also, in order to increase the algorithm efficiency, Hessian Affine, Features from Accelerated Segment Test (FAST), Hessian-blobs, and Maximally Stable ExtremalRegions (MSER) are further proposed [5-7]. In [8], Mikolajczyk et al. extract 10 different keypoint detectors within a common framework and compare them for various types of transformations. Van de Sande [9] extracts 15 types of local color features, and examines their performance on transformation invariance for image classification. Many detection methods are studied seeking a balance between keypoint repeatability and computational complexity.

After the keypoint detection, we compute a descriptor on the local patch. Feature descriptors can be divided into gradient-based descriptors, spatial frequency based descriptors, differential invariants, moment invariants, and so on. Among them, the histogram of gradient-based method has been wildly used. The gradient histogram is used to represent different local texture and shape features. The Scale Invariant Feature Transform SIFT descriptor proposed by Lowe [4] is a landmark in research of local feature descriptor. It is highly discriminative and robust to scaling, rotation, light condition change, view position change, as well as noise distortion. Since then, it has drawn considerable interests and a larger number descriptors based on the idea of SIFT emerges. SURF [10] uses the Haar wavelet to approximate the gradient SIFT operation, and uses image integral for fast computation. DAISY [11] applies the SIFT idea for dense feature extraction. The difference is that DAISY use Gaussian convolution to generate the gradient histogram. Affine SIFT [12] simulates different perspectives for feature matching, and obtains good performance on viewpoint changes, especially large viewpoint changes. Since SIFT works on the gray-scale model, many color-based SIFT descriptors are proposed to solve the color variations, such as CSIFT, RGB-SIFT, HSV-SIFT, rgSIFT, Hue-SIFT, Opponent SIFT, and Transformed-color SIFT [9, 13-14]. Most of them are obtained by computing SIFT descriptors over channels of different color space independently; therefore they usually have higher dimension (e.g. 3×128 dimension for RGB-SIFT) descriptors than SIFT. The CSIFT introduced by Albdel-Hakim et al. [13] involves a color invariant factor based on Gaussian color space model into the SIFT. It is more robust to photometrical variations. Song et al. [15] proposed compact local descriptors using an approximate affine transform between image space and color space. Burghouts et al. [16] performed an evaluation of local color invariants.

III. Methodology

A. Feature Extraction

For apparel items like dresses, the most important characteristics are their color, pattern, and shape features. In this paper we propose a novel process to fuse color feature and local pattern feature. The first is to capture the relative size of and frequency information about groups of pixels with uniform color attributes. Second the salient keypoints within the extracted color histograms are detected. Then, a local image patch around the detected feature points is computed, known as local feature descriptor. The flowchart of proposed mobile

product search system is shown in Fig. 1. The detail image processing procedures are discussed in the remainder of this section.

Color features: As an intuitive thought if two images have similar domain color or color distributions, they are regarded as matched in color, which is researched in papers [17] and [18]. Here we use RGB color space histograms to obtain the apparel items color information and evaluate the similarity between query image and database images. In the RGB color space, each pixel is represented by a combination of red, green, and blue intensities. To have the histogram not only retain enough color information but also robust to certain variations, the 256 RGB color scale is quantized into 21 bins. Besides, we adjust the histograms by weighted mean of the consecutive bins to diminish quantization problems.

Local features: In this paper we capture the local pattern features of an apparel item based on the SIFT features. The SIFT features are successfully implemented in many recognition and retrieval systems [4, 19]. It consists of four steps:

Step 1)Extrema detection: Incremental Gaussian convolution is performed on the input color histograms to create DoG space. Next, extrema are searched in three nearby scales, and the initial locations of keypoints are obtained. DoG is a convolution of a variable-scale Gaussian function $G(x,y,\sigma)$ andinput image $I(x,y)$ with regard to *x* and *y*.

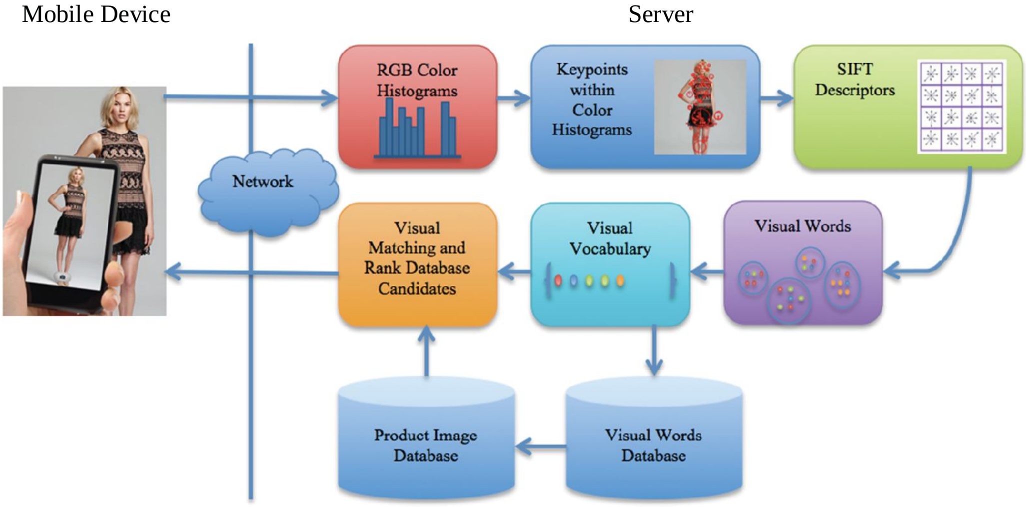


Figure 1. Product query flowchart of proposed mobile product system based on color boosted local features

$$L(x,y,\sigma)=G(x,y,\sigma)*I(x,y) \tag{1}$$

Here $L(x,y,\sigma)$ represents the scale-space of an image, and

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2} \tag{2}$$

We achieve scale invariance usingDoG.SIFT suggests that for detection of keypoint in certain scale, DoG can be obtained by doing subtraction of two images of nearby Gaussian scale-space in response to image $D(x,y,\sigma)$. Similar to LoG, keypoint can be located in location-space and scale-space using non-maximum suppression, as shown in (3).

$$D(x,y,\sigma)=\left(G(x,y,k\sigma)-G(x,y,\sigma)\right)*I(x,y)=L(x,y,k\sigma)-L(x,y,\sigma) \tag{3}$$

where*k* is a constant multiplicative factor in nearby scale-spaces. In fact, DoG and its response is an approximation to LoGand $\sigma^2\nabla^2 G$, as can be seen in the following equation.

$$G(x,y,k\sigma)-G(x,y,\sigma)\approx(k-1)\sigma^2\nabla^2 G \tag{4}$$

Step 2)Accuratekeypoint localization: Taylor expansion (up to the quadratic terms) of scale-space function $D(x,y,\sigma)$ is used and interpolated to obtain the location of keypoints, scale sub-pixel values. Low contrast points and edge response of instability points are eliminated.

Step 3)Orientation assignment: For each keypoint is assigned one or more orientations to achieve invariance to rotation. An orientation histogram is formed from the gradient orientations. The highest peak correspond

to dominant directions is detected, and any other local peak that is within 80% of the highest peak is used as auxiliary orientations to enhance the robustness of the keypoints.

Step 4) Generation of keypoint descriptor: Coordinates of each point in 16 × 16 region around the keypoint location are transformed into 4 × 4 array weighted by Gaussian window. Then multiply the weighted opposite orientation to give 8-orientation histogram, thereby obtaining 128-dimention feature descriptor.

B. Quantization

We quantize the SIFT descriptors to get the visual words using *k*-means. The SIFT descriptors are reduced in size after quantization. Every image is represented by a certain number of visual words. The computation of query is also reduced, since only the images that have common visual words with a query image will be examined, rather than compare the high dimensional SIFT descriptors of all images. However, we should notice that quantization decreases the discriminative power of SIFT descriptors, since different descriptors would be quantized into the same visual word, and hence be treated as match to each other. Then, we build the visual vocabulary using kd-tree.

C. Match and Similarity

The similarity between a query image and a database image is assessed via the extracted features. We use Euclidean distance to determine two feature descriptors match or not. The similarity between the query image and the image in the database is defined as

$$Similarity = \frac{Matched\ feature\ descritpors}{Total\ feature\ descriptors\ in\ query\ image} \tag{5}$$

D. Retrieval Performance Evaluation

In this paper, we use the normalized recall and precision [20] to evaluate the system performance. This method takes the ranking into considerations, and hence has a more comprehensive measurement of retrieval results. Recall and precision are defined as follows.

$$R_n = 1 - \frac{\sum_{i=1}^{n} R_i - \sum_{i=1}^{n} i}{(N-n)n} \tag{6}$$

and

$$P_n = 1 - \frac{\sum_{i=1}^{n} \log_{10} R_i - \sum_{i=1}^{n} \log_{10} i}{\log_{10} \frac{N!}{(N-n)!n!}} \tag{7}$$

where R_n is recall and P_n is precision; R_i is the ranking of ith relevant image in the retrieval results; n is the total number of relevant images in database; N is the total number of images in the database. The precision 1 indicates the best retrieval and 0 indicates the worst retrieval.

Since every query image will have at most 2 relevant images in our database, we consider another performance evaluation method called top-*N* retrieval rates, which evaluate whether the correct dress image (front or back) is among the top *N* returned images. We calculate the average retrieval rates at top-1, top-10, and top-20 returned images.

IV. EXPERIMENT RESULT

In this section, we compare our method with conventional color histogram, state of the art SIFT, and color SIFT features. For a fair comparison, the features of different methods are all quantized to 65 visual words to build the visual vocabulary. All experiments are conducted in an apparel dataset crawled from an online shopping website.

A. Dataset

Current product image datasets, like Stanford mobile visual search dataset [1] only contain rigid objects like cards, paintings, books, and CD/DVDs. There is no existing benchmark dataset specifically for apparel items. Hence, we collect a list of prominent brands of women apparel from Bloomingdales.com. As of October 23, 2013 it contained 1 category, 58 brands, and 3684 images. The dataset provides the library of product images as well as product brands and styles. At least two images were acquired for each item. One is the front view

on model; the other is the back view on model. Each image has a resolution of 356 × 446. The image has a gray background and a certain volume of shadow. Models in the images under similar but not exactly the same lighting conditions. The apparel item shown in the image would have occlusions and variations in viewpoint, color and shape. This dataset is practical and challenging due to the fact that it is extracted from the real online shopping department store.

B. Ranking Experiments

First, we use three different query images of the same dress to test the system performance. One is the front model image in the database, another is a model image not in the database, and the third is the dress image in front view. In the case of returning top-20 search results, the ranking of the retrieved dress images (front and back) are summarized in Table I. Figure 2 shows an example of the retrieval results.

For the model image in the database, two correct images are returns at top-10 search results with the proposed method. With RGB color histogram and original SIFT, these two images are returned within top-10 and top-20, respectively, while in top-20 Hue-SIFT can only return the front image. In practical situations a user seldom uses an exactly same image in the database as a query image. So we further test the three methods using two other typical types of image. One is a different model image not in database; the other is the dress image in front view. For the different model image not in the database, the proposed method ranks the two correct images at 1st and 3rd, and the SIFT ranks them at 10th and 14th.

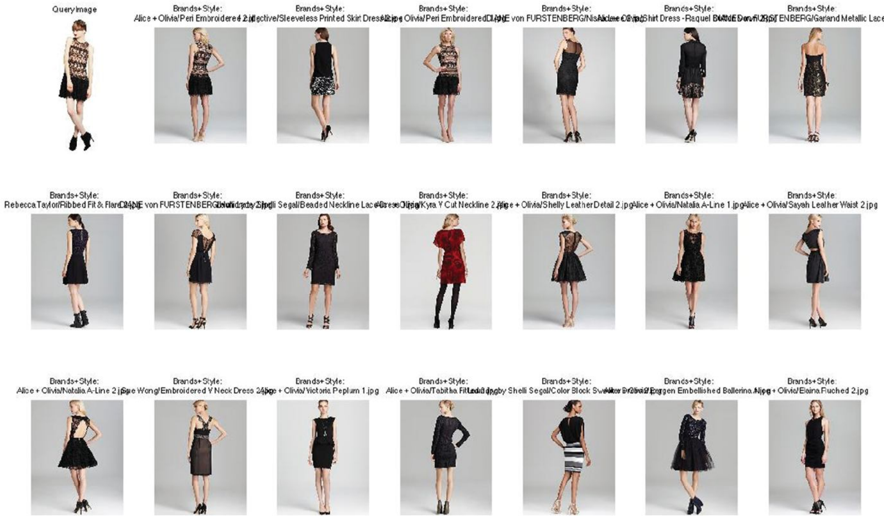
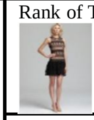


Figure 2. Retrieval results of the proposed method for different model image not in the database

TABLE I. IMAGE RETRIEVAL RANKING IN TOP-20 RESULTS

Query Image		Rank of Target Product Image - Front 				Rank of Target Product Image - Back 			
		Proposed Method	SIFT	RGB Histogram	Hue-SIFT	Proposed Method	SIFT	RGB Histogram	Hue-SIFT
Model Image in Database		1	18	1	1	8	1	9	0
Different Model Image Not in Database		3	14	0	0	1	10	0	0
Dress Image in front		5	0	0	0	0	0	0	0

The RGB color histogram and Hue-SIFT cannot rank them in top-20. For the dress image in front view, the proposed method returns the correct result at the 5th image, while all other methods cannot return them in top-20 search results. Obviously, the proposed method gains better retrieval performance thanthat of SIFT, Hue-SIFT,and RGB color histogram. The proposed color boosted feature extraction makes the colorlocal features complementeach other, and the feature descriptors are robust to a certain variations of color, shape and perspective.

C. Recall and Precision Experiments

Next, we assess our system with multiple query images. All query images are different model images that are not in the database. The recall and precision based on ranking are computed by (6) and (7), respectively. The average results are summarized in Table II. The average precision of the proposed method is higher than that of other methods. The recall is also superior to other methods. For users, if a correct image is returned after 50 ranks, such result is usually non-attractive and useless. Then, we compare the retrieval accuracies of top-1, top-10, and top-20 results in Table III. As we can see, the proposed method achieves 0.1667 of the correct retrieval at top-1 result, while other methods return none. At top-10, SIFT and RGB color histogram still cannot return any correct results. The proposed method performs 0.4583 of correct retrieval, better than 0.2083 of Hue-SIFT. At top-20, the proposed method gains 0.5000 outperforming Hue-SIFT's 0.3750 and SIFT's 0.0833. For all retrieval rates RGB color histogram can hardly have correct returns within top-20 results.

TABLE II.AVERAGE PRECISION AND RECALL OF IMAGE RETRIEVAL

Method	Proposed method	RGB color histogram	SIFT	Hue-SIFT
Recall	0.8709	0.5947	0.7500	0.7082
Precision	0.5966	0.2155	0.3065	0.3984

TABLE III. AVERAGE RETRIEVAL ACCURACIES OF TOP-1 TOP-10 AND TOP-20 RESULTS

Method	Proposed method	RGB color histogram	SIFT	Hue-SIFT
Top-1	0.1667	0.0000	0.0000	0.0000
Top-10	0.4583	0.0000	0.0000	0.2083
Top-20	0.5000	0.0000	0.0833	0.3750

V. CONTRIBUTION AND CONCLUSIONS

In this paper, we have provided a scheme for mobile product search based onthe color feature and local pattern features of apparel items. The main contribution of this work is introducing a new idea of feature extraction to address issues of existing local feature extraction methods, especially for apparel product search. Wedetect the keypoints by extracting the salient keypoints within the quantized and amended RGB color histograms, rather than SIFT, in which the keypoints are detected only on the gray density channel, or most other color SIFT methods, which perform SIFT computation over different color space channels separately. The experiment results indicate that our proposed method retains the salient color and local pattern of the apparel products whilemaintains its robustness and repeatability to certain imaging condition variation. Itoutperforms RGB color histogram, original SIFT, and Hue-SIFT.

Through observation there are several false retrievals in our system. This is mainly caused by large portion of occlusion, over complex background, great imaging condition changes like perspective and lighting, etc. Therefore, our future work includes: (i) exploring research in areas of cloth texture features, object global outline shape features andsegmentation from clustering background, as well as feature indexing to further improve retrieval performance, (ii) expanding our dataset to cover more apparel categories such as tops, tees, shorts, skirts, pants, shoes, and handbags, and (iii) extending our method to mobile video search in future work.

REFERENCES

[1] Girod, Bernd, et al. "Mobile visual search." Signal Processing Magazine, IEEE 28.4 (2011): 61-76.

[2] C. Harris, M. Stephens. A combined corner and edge detector. Proceedings of the 4th Alvey Vision Conference, Manchester, UK, 1988: 147 - 151.

[3] K. Mikolajczyk, C. Schmid. Indexing based on scale invariant interest points . Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada, 2001: 525 - 531.

[4] D. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60 (2), 91 - 110.

[5] K. Mikolajcyk and C. Schmid. “An affine invariant interest point detector.”Proceedings of the 8th InternationalConference on Computer Vision, Vancouver, Canada. 2002.

[6] J. Matas, O. Chum, M. Urban, and T. Pajdla. "Robust wide baseline stereo from maximally stable extremal regions." Proc. of British Machine Vision Conference, pages 384-396, 2002.

[7] E. Rosten, and T. Drummond. "Machine learning for high-speed corner detection". European Conference on Computer Vision 1: 430–443,2006.

[8] K. Mikolajczyk and C. Schmid, Performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Machine Intell., vol. 27, no. 10, pp. 1615–1630, 2005.

[9] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1582 -1596, 2010.

[10] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, Speeded-up robust feature, Comput. Vis. Image Understand., vol. 110, no. 3, pp. 346–359, 2008.

[11] S. Winder, G. Hua, and M. Brown, “Picking the best daisy,” in Proc. Com- puter Vision and Pattern Recognition (CVPR), Miami, FL, June 2009.

[12] JM. Morel, and G. Yu. "ASIFT: A new framework for fully affine invariant image comparison." SIAM Journal on Imaging Sciences 2.2 (2009): 438-469.

[13] A. Abdel-hakim, A. Farag. CSIFT: A SIFT descriptor with color invariant characteristics. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1978–1983, 2006.

[14] R. Behmo, P. Marcombes, A. Dalalyan, and V. Prinet, "Towards optimal naive bayes nearest neighbor," in European Conference on Computer Vision (ECCV 2010), 2010, pp. IV: 171-184.

[15] Song, Xiaohu, Damien Muselet, and Alain Trémeau. "Affine transforms between image space and color space for invariant local descriptors." Pattern Recognition, 2013.

[16] Burghouts, Gertjan J., and Jan-Mark Geusebroek. "Performance evaluation of local colour invariants." Computer Vision and Image Understanding 113.1 (2009): 48-62.

[17] H. T. Tico M and K. P. "A method of color histogram creation for image retrieval," vol. 2, June 2000, pp. 157-160.

[18] B. Zarit, B. Super, and F. Quek, "Comparison of five color models in skin pixel classification," 1999, pp. 58-63.

[19] Wang, Junqiu, HongbinZha, and Roberto Cipolla. "Combining interest points and edges for content-based image retrieval." Image Processing, 2005. ICIP 2005. IEEE International Conference on. Vol. 3. IEEE, 2005.

[20] Eakins J P, Shields K, Boardman J M. A Shape Retrieval System Based on Boundary Family Indexing. Storage and Retrieval for Image and Video Databases, 1996, 31(4): 73-80.